Файзуллаев Фирдавс

AI/ML Engineer | Архитектор LLM-систем

📧 firdavs.fayzullaev@gmail.com | 🤖 @FirdavsAIDev |

💡 Профессиональное резюме

Более 5 лет опыта в создании AI-систем, от классического ML до передовых LLM-архитектур. Прошел путь от основ data science до архитектуры корпоративных AI-решений, обрабатывающих миллионы запросов.

Ключевая экспертиза: Развертывание LLM в production (включая Llama 4, DeepSeek v3, Qwen 3), продвинутые RAG-архитектуры, мультиагентная оркестрация, real-time voice AI, оптимизация и fine-tuning моделей.

🏢 Опыт работы

AI Dynamics | AI/ML Engineer

Февраль 2024 - Октябрь 2024 | Москва (Гибрид)

Архитектор AI-систем нового поколения. Руководил переходом на Llama 4 и исследованием новых моделей (DeepSeek, Qwen) для оптимизации затрат и производительности.

- 🔹 **Advanced Multi-Modal RAG Platform (Q3 2024)**
  - **Стек:** Llama 4 (70B), DeepSeek v3, Qwen 3-VL, pgvector.
  - **Архитектура:** Гибридный поиск с re-ranking, мультимодальное понимание документов.
  - **Производительность:** 2M+ запросов/месяц, 1.5c p95 задержка, 94% точность.
- 🔹 **Autonomous Agent Network (Q2 2024)**
  - **Стек:** Claude 3.5 Sonnet, GPT-4o, LangGraph.
  - **Достижение:** Автоматизация рабочих процессов, сокращение вмешательства человека на 70%.
- 🔹 **Real-time Voice AI Platform (Q1 2024)**
  - **Стек:** Whisper large-v3 turbo, Llama 3.1 70B, custom TTS.
  - **Инновация:** Потоковая обработка с задержкой 500ms для 5000+ DAU.

TechForward Solutions | ML Engineer

Март 2022 - Январь 2024 | Москва (Удаленно)

Эволюция от традиционного ML к LLM-first подходу. Адаптация к быстрому

развитию AI от эры GPT-3 до современных мультимодельных стратегий.

- 🔹 **Enterprise Knowledge Management System (Q4 2023)**
  - **Стек:** GPT-4 Turbo, Claude 2.1, Pinecone.
  - **Влияние:** 3-кратное улучшение релевантности поиска для 500K+ документов.
- 🔹 **Financial Document Intelligence (Q2 2023)**
  - **Стек:** GPT-3.5-turbo-16k, LangChain.
  - **Достижение:** Автоматизированное извлечение данных с точностью 95%.
- 🔹 **Conversational AI Platform (Q4 2022)**
  - **Стек:** GPT-3 davinci, Redis.
  - **Масштаб:** 50K+ MAU в Telegram, WhatsApp, web.
- 🔹 **Computer Vision Pipeline (Q1 2022)**
  - **Стек:** YOLOv5, TensorFlow Serving.
  - **Применение:** Контроль качества на производстве в реальном времени с точностью 97%.

Analytics Pro | Data Scientist

Сентябрь 2020 - Февраль 2022 | Москва

Фундаментальный опыт в data science и машинном обучении.

- 🔹 **Predictive Analytics Suite (2021):** XGBoost, CatBoost. Улучшение точности прогнозов на 20%.
- 🔹 **NLP Text Classification System (2021):** BERT, Transformers. Обработка 1M+ документов в месяц.
- 🔹 **Recommendation Engine (2020):** Гибридный подход. Увеличение вовлеченности пользователей на 15%.

🚀 **Карта компетенций (AI Ecosystem 2025)**

- **Модели и API:**
  - **Облачные:** OpenAI, Anthropic, Google, Cohere.
  - **Локальные:** LLaMA, Mistral, Qwen, DeepSeek, Phi.
  - **Запуск:** Ollama, LM Studio, vLLM, TGI.
- **Фреймворки агентов:**
  - **Оркестрация:** LangChain, LangGraph, CrewAI, AutoGen.
  - **Платформы:** Flowise, Botpress.
  - **Обучение:** DSPy, RLHF, LoRA.
- **Память и RAG:**
  - **Специализированные решения: mem0** (предпочтительно).
  - **RAG-фреймворки:** LlamaIndex, LangChain.

    - **Векторные базы:** Pinecone, Weaviate, Milvus, Qdrant.
- **Инфраструктура:**
    - **Бэкенд:** FastAPI, Flask, Node.js.
    - **Оркестрация:** Docker, Kubernetes, Modal.
    - **Облака:** AWS, GCP, Azure, Hugging Face.
- **Голос и телефония:**
    - **TTS/STT:** ElevenLabs, OpenAI TTS/Whisper, Deepgram.
    - **Телефония:** Twilio, Voximplant, Asterisk.

📊 Измеримые результаты

- **Производительность систем:** Развернутые модели обрабатывают **10M+ запросов/месяц** с аптаймом 99.9%. Добился **sub-second задержки** для RAG-запросов и снижения затрат на инференс на **60%**.
- **Бизнес-результаты:** Проекты по автоматизации сэкономили компаниям более **$2M в год** и улучшили CSAT на **35%**.
- **Технические достижения:** Fine-tuning моделей Llama с улучшением производительности на 20%. Создание фреймворков для оценки 10+ моделей одновременно.

🎯 Специализированная экспертиза

- **Production LLM Systems:** Глубокое понимание оптимизации задержек, управления затратами, мониторинга и A/B-тестирования в масштабе.
- **Model Selection & Optimization:** Опыт выбора правильной модели для задачи (Llama 4 для рассуждений, DeepSeek для кода, Qwen для многоязычности).
- **RAG & Knowledge Systems:** Продвинутые стратегии поиска: гибридный поиск, re-ranking, query expansion.
- **Multi-Agent Architectures:** Создание совместных AI-систем: декомпозиция задач, протоколы связи агентов, AutoGen, LangGraph.

🎓 Образование и обучение

- **РУДН | Прикладная информатика | 2018-2022**
- **Путь непрерывного обучения:**
    - **2025:** Llama 4, DeepSeek v3, Qwen 3 multimodal.
    - **2024:** Claude 3.5, Llama 3.1 405B, LangGraph.
    - **2023:** Векторные базы, RAG-архитектуры, LangChain.

🔮 Видение и интересы

Я сфокусирован на моделях рассуждений (reasoning models), оптимизации локальных моделей для edge-устройств и мультиагентных системах. Я ищу роли,

где смогу создавать AI-системы, трансформирующие бизнес, и лидировать технические решения в AI-first продуктах.

📝 Подход к архитектуре памяти агентов

Я пришел к выводу, что специализированные слои памяти, такие как **mem0**, являются универсальным и наиболее эффективным решением. Они обеспечивают умный контекст, автоматическую суммаризацию и экономию токенов 'из коробки'. mem0 превосходит MemGPT (Letta) по производительности, экономии токенов и простоте интеграции, что подтверждается бенчмарками LOCOMO. Для чистой памяти LangGraph избыточен, но может использоваться для оркестрации логики поверх mem0.